PATENT ABSTRACTS OF JAPAN

(11)Publication number : 10-097280
(43)Date of publication of application : 14.04.1998

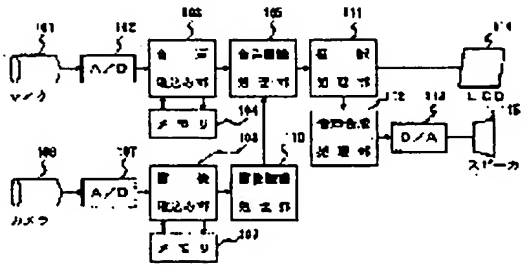
(51)Int.Cl. G10L 3/00
G10L 3/00
G10L 3/00
G06F 3/16
G06F 17/28

(21)Application number : 08-247939
(22)Date of filing : 19.09.1996

(71)Applicant : HITACHI LTD
(72)Inventor : WAKIZAKA SHINJI
KONDO KAZUO
TSUKADA TOSHIHISA
AMANO AKIO
ITO KOJI
SATO HIROKO
ISHIWATARI KAZUYOSHI

(54) SPEECH IMAGE RECOGNITION AND TRANSLATION DEVICE

(57)Abstract:
PROBLEM TO BE SOLVED: To improve the precision of translation by judging a phrase of a sentence consisting of an inputted continuous speech when features of an image of the whole mouth and its periphery corresponding to features of a speech indicate the end of the phrase of the sentence.
SOLUTION: The device consists of a microphone 101, A/D converting ICs 102 and 107, a speech input part 103, a memories 104 and 109, a speech recognizing process part 105, a camera 106, an image input part 108, an image recognizing process part 110, a translating process part 111, a speech synthesizing process part 112, a D/A converting IC 113, an LCD 114, and a speaker 115. A phrase of a sentence consisting of a continuous input speech is recognized and the words constituting the sentence are recognized to perform speech recognition and translation. Here, when features of the image of the mouth corresponding to the speech indicate the end of the phrase of the sentence, it is judged that it is the phrase of the sentence consisting of the inputted continuous speech, the character and character string in the ending of this phrase are recognized, and a mark indicating the phrase is added to accurately recognize the phrase and word ending of the speech input sentence.



LEGAL STATUS
[Date of request for examination]
[Date of sending the examiner's decision of rejection]
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]
[Patent number]
[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平10－97280

(43)公開日　平成10年(1998)４月14日

(51)Int.Cl.[6]		識別記号	FI		
Ｇ１０Ｌ	3/00	５５１	Ｇ１０Ｌ	3/00	５５１Ｃ
		５１３			５１３Ｚ
		５７１			５７１Ｇ
Ｇ０６Ｆ	3/16	３２０	Ｇ０６Ｆ	3/16	３２０Ｆ
	17/28			15/38	Ｖ

審査請求　未請求　請求項の数８　ＯＬ　（全 15 頁）　最終頁に続く

(21)出願番号　特願平8－247939

(22)出願日　平成８年(1996)９月19日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

(72)発明者　脇坂　新路
東京都小平市上水本町五丁目20番１号　株式会社日立製作所半導体事業部内

(72)発明者　近藤　和夫
東京都小平市上水本町五丁目20番１号　株式会社日立製作所半導体事業部内

(72)発明者　塚田　俊久
東京都国分寺市東恋ケ窪一丁目280番地
株式会社日立製作所中央研究所内

(74)代理人　弁理士　高田　幸彦　（外１名）

最終頁に続く

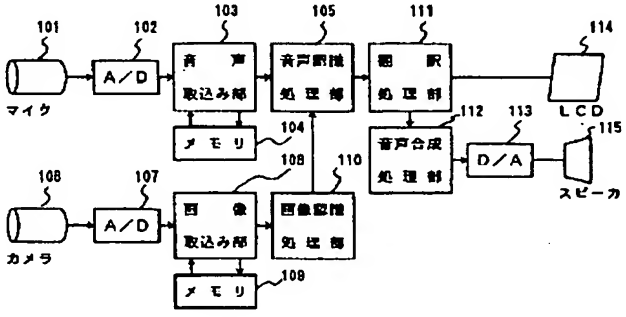
(54)【発明の名称】　音声画像認識翻訳装置

(57)【要約】

【課題】入力された一連の会話音声から成る文を高精度で翻訳する。

【解決手段】口の画像の特徴が文の文節の終わりを示した場合には、入力された一連の音声からなる該文の文節と判断すると共に該文節の終わりの文字や文字列を認識し、更に文節を示す印を付加することにより、会話音声を正確に認識して翻訳精度を高める。

図１



1

【特許請求の範囲】

【請求項１】音声と画像を取り込む手段と、取り込んだ音声データを記憶しておくメモリと、取り込んだ一連の音声に対して該音声の特徴を抽出し、音声認識処理を行う音声認識処理部と、取り込んだ画像データを記憶しておくメモリと、取り込んだ一連の画像に対して該画像の特徴を抽出し、画像認識処理を行う画像認識処理部と、音声認識及び画像認識された単語や文章の認識結果に対して翻訳したい単語や文章に翻訳する翻訳処理部とを備え、経時的に変化する音声の特徴と該音声の特徴に対応した画像の特徴の２つの相関関係から、入力された一連の音声からなる文の文節を認識し、かつ、文を構成する単語を認識することで音声認識及び翻訳する音声画像認識翻訳装置において、

音声に対応する口の画像の特徴が文の文節の終わりを示した場合には、入力された一連の音声からなる該文の文節と判断すると共に該文節の終わりの文字や文字列を認識し、更に文節を示す印を付加することを特徴とする音声画像認識翻訳装置。

【請求項２】請求項１において、同じサンプリング周波数で同時刻の間に取り込んだ一連の音声と画像から抽出した音声の特徴が、子音から母音に変化し、更に母音から無音状態に変化していく過程において、この音声の特徴に対応した口全体とその周辺近傍の画像の特徴が文の文節の終わりを示した場合には、入力された一連の音声からなる文の文節と判断することを特徴とする音声画像認識翻訳装置。

【請求項３】請求項１において、同じサンプリング周波数で同時刻の間に取り込んだ一連の日本語音声と画像から抽出した音声の特徴が、子音から母音に変化し、更に母音から無音状態に変化していく過程において、この音声の特徴に対応した口全体とその周辺近傍の画像の特徴が文の文節の終わりを示した場合には、入力された一連の音声からなる文の文節と判断すると共に該文節の終わりの助詞を認識することを特徴とする音声画像認識翻訳装置。

【請求項４】請求項１において、取り込んだ音声データを記憶しておく前記メモリを、常時、音データを取り込むメモリ（１）と、音声が入力されると該音声データを取り込むメモリ（２）とから構成すると共に、取り込んだ画像データを記憶しておく前記メモリを、常時、画像データを取り込むメモリ（３）と、音声が入力されると画像データを取り込むメモリ（４）とから構成し、

音声が入力されたと判断したときに音声データ及び画像データの取り込みを開始し、また、音声の入力が終了したと判断したときには音声データ及び画像データの取り込みを中止し、記憶した音声データ及び画像データに対して音声認識及び翻訳することを特徴とする音声画像認識翻訳装置。

【請求項５】請求項４において、音声が入力されたと判断したときに音声データ及び画像データの取り込みを開始する処理及び記憶した音声データ及び画像データの読み出し処理について、

音声の特徴が現われる任意に設定された周期（サンプリング周波数）で、常に音データを取り込んでおくメモリ（１）の該音データに対して、時間Ｔｉに取り込んだ音データの強度Ｐｉと、１つ前の周期の時間Ｔｉ-1に取り込んだ音データの強度Ｐｉ-1との差ΔＰｉの値が、任意に設定された音データの強度の境界値Ｐｔｈを超えた場合には該音データは音声データであると判断して次の音データから順にメモリ（２）に書き込むと共に、画像データの取り込みにおいては、音データが音声データであると判断した時点から画像データを常に取り込んでいるメモリ（３）への該画像データの書き込みを終了し、次の画像データから順にメモリ（４）に書き込み、メモリ（１）に記憶した音声データとメモリ（２）に記憶した音声データとを合わせて時間軸ｔ方向に順に読み出して音声波形データを形成すると共に、メモリ（３）に記憶した画像データとメモリ（４）に記憶した画像データと合わせて時間軸ｔ方向に順に読み出して一連の口の動きを示す画像データを形成することを特徴とする音声画像認識翻訳装置。

【請求項６】請求項１において、更に、音声認識及び画像認識された単語や文章の認識結果に対して認識結果の修正または補正を行う認識結果修正部を設けたことを特徴とする音声画像認識翻訳装置。

【請求項７】請求項１～６において、前記画像の特徴は、画像データの解像度変換及び２値化処理を施して正規化した人の顔画像と口全体とその周辺近傍の画像から抽出することを特徴とする音声画像認識翻訳装置。

【請求項８】請求項１～７の１項において、前記音声と画像を取り込む手段を、カメラとマイクを一体的にして構成したことを特徴とする携帯型の音声画像認識翻訳装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、携帯型音声翻訳機やＰＤＡに代表される小型情報機器や携帯型翻訳機等への応用に好適な音声画像認識翻訳装置に係り、特に、海外旅行先などで必要な会話音声を双方向に取り込んで音声認識してお互いの母国語の言語に翻訳する音声認識翻訳装置に関する。

【０００２】

【従来の技術】海外旅行者数の急激な増加に伴い、言葉の壁によるコミュニケーションの不自由さを克服するために、音声による定型会話集などの携帯型翻訳機が普及しつつある。これらの携帯型翻訳機は、予め会話に用いる文章を音声データとして登録しておき、その場の状況に合わせて必要な会話文を選択して再生する方式をとっている。したがって、この方式では、決められた定型文

章の範囲の中から選択し表現で、自分の質問や要求を相手の言葉（言語）で一方的に聞かせるのみであり、ある程度自由な言い回しで自分の言葉や相手の言葉を翻訳することができない。

【０００３】また、特開平５－３５７７６号公報（名称；「言語自動選択機能付翻訳装置」）には、マイクから入力した操作者の音声を認識して翻訳し、翻訳した言語の音声を出力するようにした携帯用の翻訳装置が開示されている。

【０００４】図１０は、このような従来の音声翻訳装置の１例を示すブロック図である。同図において、１００１は制御部、１００２は音声区間切り出し部、１００３は音声認識部、１００４は表示部、１００５は音声合成部、１００６は翻訳語データ用メモリカード、１００７は音声認識辞書部、１００８はスピーカ、１００９はマイク、１０１０はスピーカアンプ、１０１１は操作信号である。

【０００５】制御部１００１は、マイクロプロセッサを中心にして構成され、装置の各部を制御する。

【０００６】音声区間切り出し部１００２は、マイク１００９から入力された音声をデジタル信号に変換して切り出し、音声認識部１００３に送る。

【０００７】音声認識部１００３は、キーボードまたはスイッチ等により入力された操作信号１０１１を受けた制御部１００１の指示により、マイク１００９から入力して音声区間切り出し部１００２を経て切り出した音声を分析する。そして、その結果を、音声認識辞書部１００７に格納された標準音声パターンと比較して音声認識を行う。

【０００８】音声合成部１００５は、音声認識部１００３により認識した音声に対応した翻訳語を翻訳語データ用メモリカード１００６から読み込み、音声信号に変換してスピーカアンプ１０１０及びスピーカ１００８を経て音声として出力する。

【０００９】表示部１００４は、翻訳装置の使用者の指示や翻訳語の文字による表示等を行う。

【００１０】翻訳語データ用メモリカード１００６は、ＲＯＭ，フラッシュメモリ，ハードディスク等からなり、翻訳語を音声合成して出力する場合には、音声データを格納する。また、この翻訳語データ用メモリカード１００６からは、翻訳語に対応したキャラクターコードを読み込んで表示部１００４に表示する。そして、この翻訳語データ用メモリカード１００６を他の言語のものと交換することにより、複数の言語に翻訳することを可能にする。

【００１１】音声認識辞書部１００７は、ＲＯＭ，ＲＡＭ等からなり、操作者の発声に対応した標準音声パターンを格納している。この標準音声パターンは、操作者が予め登録して格納しておく。

【００１２】また、音声情報の認識に当たって、画像情報を組み合わせる技術についても、例えば、特開昭６０－１８３９９８号公報や特開昭６３－１９１１９８号公報に開示されているが、未だ正確な翻訳を可能にする具体的な認識方法を開示したものはない。

【００１３】

【発明が解決しようとする課題】前記した従来の携帯型音声翻訳装置においては、操作者の発生する音声を認識するものの、音声による定型会話集などの定型文章形式の翻訳機と同様に、予め会話に用いる文章を音声データとして登録しておき、その場の状況に合わせて必要な会話文を選択して再生する方式と機能的には変わらない。すなわち、自分の質問や要求を相手の言葉で一方的に聞かせることは可能であるが、不特定な相手の自然な会話における音声を認識して翻訳することができないことから、相手の言っていることが理解できずに会話が成立しない事態が発生するという問題がある。携帯型音声認識装置においては、自分の言いたいことを翻訳するよりは、むしろ相手の言っていることを翻訳してもらうことの方が重要である。それを実現するためには、解決しなければならない難しい技術課題が沢山ある。

【００１４】海外旅行先などで交わされる会話のような、比較的短い文章から成る連続音声を認識して翻訳するためには、連続音声を構成している単語と単語、例えば、日本語の場合は、名詞，動詞，助詞などの区別や、文の区切りをはっきりと認識しなければ正しい翻訳を行なうことができない。そこで、音声の入力を文節単位に区切って入力することが考えられるが、余り間を置き過ぎても不自然な会話文になってしまう。すなわち、海外旅行先などで交わされる会話のような、比較的短い文章から成る連続音声の入力であって多少文節を意識する程度の連続音声の入力により、音声を認識して翻訳しなければならない。

【００１５】そこで、本発明の目的は、少しでも会話らしい相互の会話音声認識翻訳を可能にする携帯型音声認識翻訳機を実現するために、入力された一連の音声から成る文の文節を認識し、また、文を構成する単語を認識することで会話音声を認識して正しく翻訳することができる音声認識翻訳装置を提供することにある。

【００１６】

【課題を解決するための手段】前記した目的を達成するために、本発明になる音声認識翻訳装置は、音声と画像を取り込む手段と、取り込んだ音声データを記憶しておくメモリと、取り込んだ一連の音声に対して該音声の特徴を抽出し、音声認識処理を行う音声認識処理部と、取り込んだ画像データを記憶しておくメモリと、取り込んだ一連の画像に対して該画像の特徴を抽出し、画像認識処理を行う画像認識処理部と、音声認識及び画像認識された単語や文章の認識結果に対して翻訳したい単語や文章に翻訳する翻訳処理部とを備え、経時的に変化する音声の特徴と該音声の特徴に対応した画像の特徴の２つの

5

相関関係から、入力された一連の音声からなる文の文節を認識し、かつ、文を構成する単語を認識することで音声認識及び翻訳する音声画像認識翻訳装置において、音声に対応する口の画像の特徴が文の文節の終わりを示した場合には、入力された一連の音声からなる該文の文節と判断すると共に該文節の終わりの文字や文字列を認識し、更に文節を示す印を付加することを特徴とする。このようにすれば、音声入力文の文節と語尾（日本語においては助詞）を正確に認識することができるので、翻訳の精度が向上する。

【００１７】具体的には、同じサンプリング周波数で同時刻の間に取り込んだ一連の音声と画像から抽出した音声の特徴が、子音から母音に変化し、更に母音から無音状態に変化していく過程において、この音声の特徴に対応した口全体とその周辺近傍の画像の特徴が文の文節の終わりを示した場合には、入力された一連の音声からなる文の文節と判断するようにする。

【００１８】また、同じサンプリング周波数で同時刻の間に取り込んだ一連の日本語音声と画像から抽出した音声の特徴が、子音から母音に変化し、更に母音から無音状態に変化していく過程において、この音声の特徴に対応した口全体とその周辺近傍の画像の特徴が文の文節の終わりを示した場合には、入力された一連の音声からなる文の文節と判断すると共に該文節の終わりの助詞を認識するようにする。

【００１９】また、取り込んだ音声データを記憶しておく前記メモリを、常時、音データを取り込むメモリ（１）と、音声が入力されると該音声データを取り込むメモリ（２）とから構成すると共に、取り込んだ画像データを記憶しておく前記メモリを、常時、画像データを取り込むメモリ（３）と、音声が入力されると画像データを取り込むメモリ（４）とから構成し、音声が入力されたと判断したときに音声データ及び画像データの取り込みを開始し、また、音声の入力が終了したと判断したときには音声データ及び画像データの取り込みを中止し、記憶した音声データ及び画像データに対して音声認識及び翻訳するようにする。

【００２０】また、音声が入力されたと判断したときに音声データ及び画像データの取り込みを開始する処理及び記憶した音声データ及び画像データの読み出し処理について、音声の特徴が現われる任意に設定された周期（サンプリング周波数）で、常に音データを取り込んでおくメモリ（１）の該音データに対して、時間Ｔｉに取り込んだ音データの強度Ｐｉと、１つ前の周期の時間Ｔｉ-1に取り込んだ音データの強度Ｐｉ-1との差ΔＰｉの値が、任意に設定された音データの強度の境界値Ｐｔｈを超えた場合には該音データは音声データであると判断して次の音データから順にメモリ（２）に書き込むと共に、画像データの取り込みにおいては、音データが音声データであると判断した時点から画像データを常に取り込んでいるメモリ（３）への該画像データの書き込みを終了し、次の画像データから順にメモリ（４）に書き込み、メモリ（１）に記憶した音声データとメモリ（２）に記憶した音声データとを合わせて時間軸ｔ方向に順に読み出して音声波形データを形成すると共に、メモリ（３）に記憶した画像データとメモリ（４）に記憶した画像データと合わせて時間軸ｔ方向に順に読み出して一連の口の動きを示す画像データを形成するようにする。

6

【００２１】また、更に、音声認識及び画像認識された単語や文章の認識結果に対して認識結果の修正または補正を行う認識結果修正部を設ける。

【００２２】また、前記画像の特徴は、画像データの解像度変換及び２値化処理を施して正規化した人の顔画像と口全体とその周辺近傍の画像から抽出する。

【００２３】また、前記音声と画像を取り込む手段を、カメラとマイクを一体的にして構成して携帯型の音声画像認識翻訳装置を構成する。

【００２４】上述のような構成の音声認識翻訳装置を用いることにより、少しでも会話らしい相互の音声認識翻訳を可能にする携帯型音声認識翻訳機が得られる。

【００２５】

【発明の実施の形態】以下、図面を参照しながら、本発明の実施の形態について詳細に説明する。

【００２６】図１は、本発明の一実施形態に係る音声画像認識翻訳装置の構成を示すブロック図である。この図１に示した音声画像認識翻訳装置は、携帯型音声認識翻訳機である。このような装置は、ＣＰＵやメモリや専用ＩＣ等のいくつかのＬＳＩで構成することができる。また、チップとして、半導体素子上に構成することもできる。

【００２７】図１において、１０１は音声を取り込むための指向性マイクであり、例えば、海外旅行先の空港や駅構内，飛行機内，ホテル，観光地，レストランやショッピング等で交わされる会話音声を取り込む。

【００２８】１０２は１６ビットのアナログ／デジタル（Ａ／Ｄ）変換ＩＣであり、前記マイク１０１内のフィルタやアンプにより音声帯域以外の音が取り除かれて雑音処理された音声データの連続的なアナログ信号を、音声のサンプリング周波数，例えば１２ｋＨｚでサンプリングしてデジタル信号に変換する。

【００２９】１０３は音声取り込み部であり，Ａ／Ｄ変換ＩＣ１０２でサンプリングされた１６ビットの音声データに対して、シリアルデータからパラレルデータにシリアル／パラレル変換を行ってレジスタ等に一旦格納しておくためのものである。

【００３０】１０４は、前記音声取り込み部１０３により取り込んだ音声データ，例えば，会話音声の１フレーズ分の連続音声データを記憶しておくためのメモリであり，また，連続音声データを書き込めるだけの必要最小限の容量を持つメモリである。連続音声データのメモリ

の書き込みは、ＣＰＵ等のソフトウエア処理で行っても、専用のハードウニアで行っても良い。

【００３１】１０５は音声認識処理部であり、メモリ１０４に書き込まれた連続音声データに対して、デジタルフィルタ，音声分析，音声区間検出，照合，判定等の一連の音声認識処理を行う。ここで、音声認識に必要な音響モデルデータ，辞書データ，文法データは、この音声認識処理部１０５内においてメモリ等に登録し格納しておく。音声認識処理は、ＣＰＵ，ＤＳＰ等のソフトウニア処理で行っても、専用のハードウエアで行っても良い。

【００３２】１０６は、画像を取り込むための高解像度カメラである。これは、例えば、ＣＣＤカメラであり、海外旅行先の空港や駅構内，飛行機内，ホテル，観光地，レストランやショッピング等で交わされる会話音声に合わせて該音声を発生する人の口の動きを画像データとして取り込む。

【００３３】１０７は１６ビットのアナログ／デジタル（Ａ／Ｄ）変換ＩＣであり、前記ＣＣＤカメラ１０６からのアナログ信号を、音声のサンプリング周波数に同期して、例えば、１２ｋＨｚでサンプリングしてデジタル信号に変換する。

【００３４】１０８は画像取り込み部であり、前記Ａ／Ｄ変換ＩＣ１０７でサンプリングした１６ビットの画像データに対して、シリアルデータからパラレルデータにシリアル／パラレル変換を行ってレジスタ等に一旦格納しておくためのものである。

【００３５】１０９は、画像取り込み部１０８により取り込んだ画像データ、例えば、会話音声の１フレーズ分の連続画像データを記憶しておくためのメモリであり、また、連続画像データを書き込めるだけの必要最小限の容量を持つメモリである。連続画像データのメモリの書き込みは、ＣＰＵ等のソフトウエア処理で行っても専用のハードウエアで行っても良い。

【００３６】１１０は画像認識処理部であり、前記メモリ１０９に書き込まれた連続画像データに対して、デジタルフィルタ，画像変換，２値化処理，画像解析，特徴抽出，照合，判定等の一連の画像認識処理を行う。ここで、画像認識に必要な画像モデルデータ，辞書データ，文法データは、画像認識処理部１０５内において、メモリ等に登録して格納しておく。画像認識処理は、ＣＰＵやＤＳＰ等のソフトウエア処理で行っても、専用のハードウエアで行っても良い。ここで、画像認識処理した結果は、音声認識処理部１０５に渡す。

【００３７】１１１は、前記音声認識処理部１０５から出力された会話音声の認識結果に対して翻訳したい言語に翻訳処理を行う翻訳処理部である。音声認識処理部１０５から出力する認識結果は、例えば、日本語であれば名詞，助詞，動詞，副詞等のかな漢字まじりのテキスト文章である。翻訳処理では、これらのかな漢字まじり文

章に対して、構文解析及び辞書、文法規則、事例等からの文章生成を行い、翻訳結果を出力する。

【００３８】１１２は、前記翻訳処理部１１１から出力された翻訳結果を、会話文に適した音声に変換して音声出力する音声合成処理部である。この音声合成処理部１１２では、より自然な会話文音声にするために、文章を構成している単語の発音やアクセント、更に、文章全体の抑揚を最適化して会話文の音声合成を行い、相手側に対して聞き取りやすい自然な音声を出力するための処理も行う。

【００３９】１１３は１６ビットのデジタル／アナログ（Ｄ／Ａ）変換ＩＣであり、前記音声合成処理部１１２から出力された音声のデジタル信号を、例えば、ローパスフィルタ（ＬＰＦ）を経由して音声周波数帯域１２ｋＨｚでアナログ信号に変換する。

【００４０】１１４は、音声認識結果の途中経過や翻訳結果をテキストで表示するための液晶ディスプレイ（ＬＣＤ）である。

【００４１】１１５は、音声認識結果やその途中経過，翻訳結果を音声合成して音声出力するためのスピーカである。

【００４２】図２は、音声認識をサポートするためのアプローチとして、音声を発生する人の口の動きの画像を取り込み、画像認識して音声認識と共に音声の内容を解読する方法の一例を説明するための図である。

【００４３】図１に示したカメラ１０６，例えば、ＣＣＤカメラにより、音声入力に同期して音声を発生する人の口の動きの画像を取り込む。

【００４４】２０１は、音声入力と同期して取り込んだ任意の時点での会話音声を発生している人の静止画像である。画像の解像度は、縦ｍドット×横ｎドット×深さ１ビットである。この画像２０１は、音声入力に同期して、例えば、音声のサンプリング周波数を１２ｋＨｚに設定し、この音声の特徴を抽出するためのフレームを音声のサンプリング点数を２４０ポイントにした場合は、２０ｍｓのフレーム単位で音声入力に同期させて取り込むことになる。音声の取り込みにおける２０ｍｓのフレーム単位を画像の取り込みに適用すると、２０ｍｓで１フレームの画像を取り込むことになるので、１秒間に５０フレームの画像を処理することになる。これは、現行のＮＴＳＣ等の動画像である３０フレーム／秒よりも時間軸の分解能が高くなる。

【００４５】２０２は、取り込んだ画像２０１に対して、カラー画像から濃淡画像に変換する濃淡化処理を行い、その後、２値化画像を得る２値化処理を施し、口の形の特徴を抽出するために取り出した画像である。

【００４６】２０３は、画像２０２に対して、正規化，エッジ検出，平滑化処理を施し、口の輪郭等の特徴を抽出した画像である。この画像２０３の解像度は、画像の特徴を十分に表すことができる必要最小限の解像度とす

る。

【００４７】図３は、音声及び画像を取り込むためのカメラを内蔵したマイクの構成を説明するための図である。図１に示した音声画像認識翻訳装置において、マイク１０１とカメラ１０６を一体化したものである。特に、携帯型音声認識翻訳機のような装置においては、部品点数の削減，低コスト化，低価格化が重要である。また、図２を参照して説明したように、口の動きの画像認識を行うことから、音声を取り込むマイクにマイクを内蔵すれば、このマイクに向かって音声を発生する口の画像を比較的容易に取り込むことができる。

【００４８】３０１は、音声及び画像を取り込むためのカメラを内蔵したマイク本体である。外形は、円筒形であっても、角形であっても良い。３０２はマイクであり、コンデンサ型マイクや抵抗型マイク等で構成される。３０３はレンズであり、音声を発生する人物の顔や口の画像が取り込めるようにチューニングしておく。３０４はＣＣＤカメラであり、レンズ３０３から進入してきた音声を発生する人物の顔や口の画像を取り込む。

【００４９】３０５は音声及び画像データを伝送するためのケーブルであり、音声信号ケーブル３０６と画像信号ケーブル３０７とを備える。図１に示した音声画像認識翻訳装置おいては、音声信号ケーブル３０６はＡ／Ｄ変換ＩＣ１０２に接続し、画像信号ケーブル３０７はＡ／Ｄ変換ＩＣ１０７に接続する。

【００５０】図４は、このような本発明になる音声画像認識翻訳装置において、実際に、会話音声による会話文が、音声認識と画像認識とから認識され、翻訳されるまでを説明するための図である。

【００５１】入力された会話音声による会話文例の内容は、「コウエンハ　ドコデスカ」である。入力される音声の発生スピードは、会話における自然なスピードであり、認識率を高めるために丁寧にはっきりと発声している。

【００５２】４０１は、入力された会話音声による会話文例における「コウエンハ　ドコデスカ」の音声波形である。時間軸ｔにおける音声波形の振幅は、音声強度を表している。この音声波形に対して、音声のサンプリング周波数を１２ｋＨｚに設定し、音声の特徴を抽出するためのフレームを音声のサンプリング点数を２４０ポイントにした場合は、２０ｍｓのフレーム単位で音声を取り込むことになる。

【００５３】４０２～４０４は、音声の特徴を抽出するためのフレームを音声のサンプリング点数を２４０ポイントにした場合の２０ｍｓ単位の音声特徴フレームである。音声の特徴は、一般的に、音声認識で採用されている音声分析により抽出された特徴である。ここで、音声特徴フレーム４０２で表される特徴は、「コウエンハ」の動的な音声の特徴である。また、音声特徴フレーム４０３で表される特徴は、「ハ　　」の静的な音声の特徴である。再び、音声特徴フレーム４０４で表される特徴は、「ドコデスカ」の動的な音声の特徴である。

【００５４】このような音声の特徴抽出から音声認識を行うことは可能であるが、音声の始まり（語頭）や終わり（語尾）において誤認識が起こりやすい。例えば、「コウエンハ」が「コウエンヘ」になったり、「ドコデスカ」が「ココデスカ」になったりする。また、この音声認識翻訳装置を海外旅行先などで活用することを考えると、周囲の雑音などにより、更に認識率が低下する。一方、翻訳においては、音声認識の結果次第で、翻訳精度が変化する。特に、日本語から英語に翻訳する例をとってみると、語頭や語尾において誤認識が起こって、「コウエンハ」が「コウエンヘ」になったり、「ドコデスカ」が「ココデスカ」になったりすると、正しく翻訳できなくなる。また、文節も認識できないと正しい翻訳は困難になる。　そこで、翻訳のために、音声認識の精度を高める手段として、音声の始まり（語頭）や終わり（語尾）における口の動きに着目した。４０５～４０７は、音声特徴フレーム４０２～４０４で示した音声の特徴を抽出するためのフレームを音声のサンプリング点数を２４０ポイントにした場合の２０ｍｓ単位の音声特徴フレームに対応させて、口の動きの画像を取り込み、図２を参照して説明したような処理により特徴を抽出した画像である。ここで、画像４０５で表される特徴は、「コウエンハ」の動的な口の形の連続的な動きによる画像の特徴である。また、画像４０６で表される特徴は、「ハ　　」の静的な口の形の画像の特徴である。再び、画像４０７で表される特徴は、「ドコデスカ」の動的な口の形の連続的な動きによる画像の特徴である。したがって、会話文のような連続音声の認識に対して、口の動きの画像認識を行うことにより，音声の始まりや終わりにおける文章の認識精度を高め、また、文節の認識を行うことにより、翻訳精度を高めることができるようにした。同じサンプリング周波数で同時刻の間に取り込んだ一連の音声と画像の特徴を抽出し、抽出した音声の特徴が子音から母音に変化し、更に母音から無音状態に変化していく過程において、この音声の特徴に対応した口全体とその周辺近傍の画像の特徴が、文の文節の終わりを示した場合には、入力された一連の音声からなる文の文節と判断し、文節の終わりの文字や文字列を認識するようにする。日本語の場合には、文節の終わりの助詞を認識するようにすることが好ましい。更に、例えば、「コウエンハ」の動的な口の形の連続的な動きによる画像の特徴からも画像認識を行い、音声認識と共に、「公園は」を認識する。あるいは、音声認識がなされなかった場合でも、画像の認識だけから「公園は」を認識するようにする。これにより、ロバストネス向上が期待できる。

【００５５】４０８～４１２は、入力された会話音声による会話文例「コウエンハ　ドコデスカ」の音声認識結

果である。結果は、かな漢字まじりのテキスト文章で出力する。ここで、文章を構成している単語や文節を認識することで、「公園」４０８と「は」４１０との間にスラッシュ４０９を挿入し、「は」４１０と「どこですか」４１２との間にスラッシュ４１１を挿入する。これにより、翻訳処理を容易にし、翻訳精度を高めることができるようになる。

【００５６】４１３は、認識結果「公園／は／どこですか」の翻訳結果であり「Ｗｈｅｒｅｉｓ　ｔｈｅ　ｐａｒｋ？」をテキストで出力する。そして、この翻訳結果４１３は、音声合成して音声出力する。

【００５７】図５は、このような本発明になる音声画像認識翻訳装置において、音声の取り込み及び画像の取り込み方法を説明するための図である。

【００５８】図５の（ａ）は、発明者が実際に音声として発生した「山田」の音声の波形である。図５の（ｂ）は、音声及び画像を取り込むためのメモリである。

【００５９】図５（ａ）において、図の横軸方向（図における右方向）は時間ｔを表し、縦軸方向（図における上下方向）は音声波形の振幅強度を表している。

【００６０】区間５０１に示す波形は、常時、音声を取り込んでいる状態における音声の波形である。ここで、区域１に示す波形は、無音状態の時の波形であり、区域２に示す波形は、無音状態から音声が始まる最初の音声波形を示している。音声が発生されていない場合は、区域１で示すような波形が連続するので、区域２でも区域１と同様な波形が現われる。そこで、例えば、連続的に観測される音の波形において、ある時間領域だけ音の波形データをメモリに記録しておき、常に新しい音データをメモリに格納し、古い音データから消していくような必要最小限のメモリ容量を持つメモリを考える。例えば、０．１秒程度の音声データが格納できる容量を持たせる。

【００６１】図５（ｂ）におけるメモリ（１）５０６は、連続的に観測される音の波形において、ある時間領域だけ音の波形データをメモリに記録しておき、常に新しい音データをメモリに格納し、古い音データから消していくために必要最小限のメモリ容量ｎを持つメモリである。５０７は、メモリ（１）５０６におけるライトアドレスＷＡである。５０８は、メモリ（１）５０６におけるリードアドレスＲＡである。音声が取り込まれると、ライトアドレスＷＡ５０７の示すアドレスに音声データをライトし、ライトアドレスＷＡ５０７をインクリメントしておく。この動作をメモリ（１）５０６の先頭アドレスから順に繰り返し、アドレスの最後までライトしたならば再び先頭アドレスに戻って同様に処理を繰り返す。例えば、図５（ａ）の区間５０１において、無音状態の区域１の波形データは、図５（ｂ）におけるメモリ（１）５０６の領域１に書き込まれる。また、図５（ａ）の区間５０１において、無音状態から音声が始まる区域２の最初の音声波形データは、図５（ｂ）におけるメモリ（１）の領域２に書き込まれる。区間５０１において、区域２のデータが区域１のデータよりも前に示されているのは、メモリ（１）５０６において、領域２のデータが領域１のデータよりも新しいことを表すためである。したがって、音声認識するためには、区間５０２に示す波形データのように、領域１のデータを領域２のデータよりも先に読み出すことになる。これにより、音データを常に取り込むことが可能になる。メモリ（１）５０６の音データリードに関しては後で述べる。

【００６２】次に、無音状態から、音声が始まる最初の音声波形が観測されたならば、次から取り込まれる音データは音声であると判断し、別なメモリに格納することを考える。図５（ａ）における区間５０３が音声データである。これにより、メモリ容量を大幅に削減できる。

【００６３】図５（ｂ）におけるメモリ（２）５０９は、無音状態から、音声が始まる最初の音声波形が観測されたならば、次から取り込まれる音データは音声であると判断し、その音声データを格納するためのメモリである。５１０は、メモリ（２）５０９におけるライトアドレスＷＡである。５１１は、メモリ（２）５０９におけるリードアドレスＲＡである。メモリ容量は、例えば、単語程度の認識であれば、３～５秒の音声データを格納することができる容量を持たせる。メモリ（２）における３の領域には、図５（ａ）における区間５０３の音声データ３が格納される。ここで、音声データが、音声状態から再び無音状態になり、無音と判断したならば、メモリ（２）５０９への音声データの書き込みを中止しても良い。以上に説明した音声の取り込み処理により、図５（ａ）に示す区間５０２，５０３の一連の音声波形データ１，２，３は、図５（ｂ）に示すメモリ（１）５０６及びメモリ（２）５０９に書き込まれ、これらのメモリ（１）及び（２）に書き込まれた音声波形データは、５１２に示すように時系列に読み出すことができる。

【００６４】図５（ｂ）におけるメモリ（３）５１３は、メモリ（１）５０６と同様に、連続的に観測される音の波形に対応させて、ある時間領域だけ口の動きの画像データをメモリに記録しておき、常に新しい画像データをメモリに格納し、古い画像データから消していくために必要最小限のメモリ容量ｍを持つメモリである。５１４は、メモリ（３）５１３におけるライトアドレスＷＡである。５１５は、メモリ（３）５１３におけるリードアドレスＲＡである。画像が取り込まれると、ライトアドレスＷＡ５１４の示すアドレスに画像データをライトし、ライトアドレスＷＡ５１４をインクリメントしておく。この動作をメモリ（３）５１３の先頭アドレスから順に繰り返し、アドレスの最後までライトしたならば再び先頭アドレスに戻って同様の処理を繰り返す。例えば、図５（ａ）の区間５０１において、無音状態の区域

1の波形データに対応した口の動きの画像データは、図5（b）におけるメモリ（3）513の領域1に書き込まれる。

【0065】また、図5（a）の区間501において、無音状態から音声が始まる区域2の最初の音声波形データに対応した口の動きの画像データは、図5（b）におけるメモリ（3）513の領域2に書き込まれている。区間501において、区域2のデータが区域1のデータよりも前に示されているのは、メモリ（3）513において、領域2のデータが領域1のデータよりも新しいことを表すためである。したがって、音声認識に対応した画像認識を行うためには、画像においても、区間502に示す波形データのように、メモリ（3）513において、領域1のデータを領域2のデータよりも先に読み出すことになる。これにより、音データに対応した画像データを常に取り込むことが可能になる。

【0066】次に、無音状態から、音声が始まる最初の音声波形が観測されたならば、次から取り込まれる音データは音声であると判断し、画像データにおいても同様に、別なメモリに格納することを考える。図5（a）における区間503が音声データであり、この音声データ対応した動的な口の動きを表す画像データが存在する。これにより、メモリ容量を大幅に削減できる。

【0067】図5（b）におけるメモリ（4）516は、無音状態から、音声が始まる最初の音声波形が観測されたならば、次から取り込まれる音データは音声であると判断し、画像データにおいても同様に、その音声データに対応した画像データを格納するためのメモリである。517は、メモリ（4）516におけるライトアドレスWAである。518は、メモリ（4）516におけるリードアドレスRAである。このメモリ（4）516における領域3には、図5（a）における区間503の音声データ3に対応する画像データが格納される。ここで、音声データが、音声状態から再び無音状態になり、無音と判断したならば、メモリ（4）516への画像データの書き込みを中止しても良い。

【0068】以上の説明した画像の取り込み処理により、図5（a）に示す区間502，503の一連の音声波形データ1，2，3に対応した画像データは、図5（b）に示すメモリ（3）513及びメモリ（4）516に書き込まれ、これらのメモリ（3），（4）に書き込まれたデータは、519に示すように時系列に読み出すことができる。

【0069】図6は、図5における音声及び画像の取り込み処理を示すフローチャートである。

【0070】処理ステップ601では、常に音データを取り込んでメモリ（1）にライトし、常に画像データを取り込んでメモリ（3）にライトする。例えば、音データは、12kHzでサンプリングされた音声を含む音データである。また、画像データは、音声の取り込みに同期してサンプリングされた口の動きを含む人物の顔画像データである。

【0071】処理ステップ602では、メモリ（1）にライトされた音データに対して、音データの振幅強度Pをサンプリングされるデータ毎に観測し、その振幅強度Piが任意に設定されたスレッシュホールドレベルPthを超えなかった場合は、ライトアドレスWA507，514をインクリメントし、処理ステップ601の処理を繰り返すように戻る。また、この処理ステップ602では、メモリ（1）にライトした音データに対して、音データの振幅強度Pをサンプリングされるデータ毎に観測し、その振幅強度Piが任意に設定されたスレッシュホールドレベルPthを超えた場合は、その音データは、音声であると判断する。

【0072】処理ステップ603では、処理ステップ602で音データが音声であると判断したときに、メモリ（1）506及びメモリ（3）513のライトアドレスWA507，514を記憶する。

【0073】処理ステップ604では、次にサンプリングされる音データは音声データであると判断し、メモリ（2）509へは音声データPi+1からライトし、メモリ（4）516へは画像データIi+1からライトする。以上により、音声認識に必要な音声データ及び画像データを取り込むことができる。

【0074】処理ステップ605では、音声認識するために、メモリ（1）506及びメモリ（2）509に書き込まれている音声データと、メモリ（3）513及びメモリ（4）516に書き込まれている画像データを読み出すにあたり、処理ステップ603で記憶しておいたメモリ（1）506のライトアドレスWA507の次のアドレスを該メモリ（1）506のリードアドレスの先頭アドレスRA（＝WA＋1）508として該メモリ（1）506に格納されている音声データを総て読み出す。また、処理ステップ603で記憶しておいたメモリ（3）513のライトアドレスWA514の次のアドレスを該メモリ（3）513のリードアドレスの先頭アドレスRA（＝WA＋1）515として該メモリ（3）513に格納されている画像データを総て読み出す。

【0075】最後に、処理ステップ606では、大半の音声データが格納されているメモリ（2）509に対して、リードアドレスの先頭アドレスRA（先頭）511から該メモリ（2）509に格納されている音声データを総て読み出す。また、大半の画像データが格納されているメモリ4（516）に対して、リードアドレスの先頭アドレスRA（先頭）518から該メモリ（4）516に格納されている画像データを総て読み出す。

【0076】図7は、常時、音データを取り込み、そのデータが無音状態であるか、あるいは、音声であるかを判断する場合のやり方を例示している。図5（a）において、波形504は、音声と判断された最初に取り込ま

れた波形であり、５０５は、音声であると判断された時点で発生されるトリガ（フラグ）あるいは信号である。この音声と判断された最初に取り込まれた波形を詳細に見てみると、図７に波形７０１で示すような波形となっている。ここで、７０２は、時刻ｔ（ｉ）においてサンプリングされた音声データＰｉである。７０３は、時刻ｔ（ｉ－１）においてサンプリングされた音声データＰi-1 である。

【００７７】例えば、音声データＰｉ７０２と音声データＰi-1 ７０３の差分値ΔＰを観測し、任意に設定されたスレッシュホールドレベルＰｔｈと比較して図６における処理６０２を行ってもよい。また、差分値ΔＰの積分値ΣΔＰを観測し、任意に設定されたスレッシュホールドレベルＰｔｈと比較して、図６における処理６０２を行ってもよい。更に、差分値ΔＰをある時間内に複数回（ｋ回）観測することで、その観測回数を任意に設定された回数（スレッシュホールド回数）と比較して、図６における処理６０２を行ってもよい。

【００７８】図８は、上述した携帯型音声画像翻訳機のイメージ及び外観の例を示す図である。図８（ａ）は、本発明になる携帯型音声翻訳機を海外旅行者が使用している場面を示している。ユーザである海外旅行者は、携帯型翻訳機のディスプレイ及び音声入出力手段を介して、例えば、ショッピングにおいて店員と会話をする際に、自分の話す内容を相手のわかる言葉に翻訳し、意図を伝え、逆に、相手の言っている言葉を自分のわかる言葉に翻訳し、相手の意図を理解する。特に、会話における日本語，英語，ドイツ語，フランス語，イタリア語，ロシア語，中国語等の言語については、各国の言語に対応することができ、限定されることはない。　図８（ｂ）は、携帯型翻訳機の外観図であり、８０１は、携帯型翻訳機の本体である。

【００７９】８０２は、多方向性マイクであって、空港や駅構内，飛行機内，バスや地下鉄やタクシー等の乗り物車内，観光地建物内等での会話音声に含まれる各場所での雑音を除去する目的で使用され、会話音声が無いときには各場所での全体音を取り込む。

【００８０】８０３は、図３に示したようなＣＣＤカメラ内蔵マイクで指向性があり、海外旅行先での空港や駅構内，飛行機内，ホテル，観光地，レストランやショッピング等で交わされる会話音声及び口の動きを含む顔の画像をアナログ信号として取り込む。

【００８１】８０４は、音声出力手段であり、音声認識により翻訳した内容を報音するためのスピーカやイヤホンからなる。

【００８２】８０５は、音声認識した結果やその修正，補正結果及び翻訳結果の内容を表示するためのディスプレイである。

【００８３】８０６は、ＩＣカードで、例えば、日本語から中国語に音声認識翻訳するための音響モデル，単語辞書，文法辞書，翻訳事例辞書，音声合成用の音声辞書等をメモリやハードディスクに格納して搭載している。

【００８４】８０７は、ＩＣカードで、例えば、中国語から日本語に音声認識翻訳するための音響モデル，単語辞書，文法辞書，翻訳事例辞書，音声合成用の音声辞書等をメモリやハードディスクに格納して搭載している。

【００８５】図９は、本発明に係る音声画像認識翻訳装置の他の実施形態の構成を示すブロック図である。例えば、この図９に示した音声画像認識翻訳装置は、携帯型音声認識翻訳機であり、ＣＰＵやメモリや専用ＩＣ等のいくつかのＬＳＩで構成される装置であっても、半導体素子上に構成されるチップであっても良い。

【００８６】図９において、９０１は音声を取り込むための指向性マイクであり、例えば、海外旅行先の空港や駅構内，飛行機内，ホテル，観光地，レストランやショッピング等で交わされる会話音声を取り込む。

【００８７】９０２は１６ビットのアナログ／デジタル（Ａ／Ｄ）変換ＩＣであり、マイク９０１内のフィルタやアンプにより音声帯域以外の音が取り除かれ、雑音処理された音声データの連続的なアナログ信号を、音声のサンプリング周波数，例えば１２ｋＨｚでサンプリングしてデジタル信号に変換する。

【００８８】９０３は音声取り込み部であり、前記Ａ／Ｄ変換ＩＣ１０２でサンプリングされた１６ビットの音声データに対してシリアルデータからパラレルデータにシリアル／パラレル変換を行ってレジスタ等に一旦格納しておくためのものである。

【００８９】９０４は、前記音声取り込み部９０３により取り込んだ音声データ、例えば、会話音声の１フレーズ分の連続音声データを記憶しておくためのメモリであり、また、連続音声データを書き込めるだけの必要最小限の容量を持つメモリである。連続音声データのメモリの書き込みは、ＣＰＵ等のソフトウエア処理で行っても、専用のハードウエアで行っても良い。

【００９０】９０５は、音声認識処理部であり、メモリ９０４に書き込まれた連続音声データに対して、デジタルフィルタ，音声分析，音声区間検出，照合，判定等の一連の音声認識処理を行う。ここで、音声認識に必要な音響モデルデータ，辞書データ，文法データは、この音声認識処理部９０５内において、メモリ等に登録し格納しておく。音声認識処理は、ＣＰＵやＤＳＰ等のソフトウエア処理で行っても、専用のハードウエアで行っても良い。

【００９１】９０６は、画像を取り込むための高解像度カメラであり、例えば、ＣＣＤカメラである。この高解像度カメラ９０６は、海外旅行先の空港や駅構内，飛行機内，ホテル，観光地，レストランやショッピング等で交わされる会話の音声に合わせて、この音声を発生する人の口の動きを画像データとして取り込む。

【００９２】９０７は１６ビットのアナログ／デジタル

17

（A／D）変換ICであり、CCDカメラ906からのアナログ信号を、音声のサンプリング周波数に同期して、例えば、12kHzでサンプリングしてデジタル信号に変換する。

【0093】908は画像読み取り部であり、前記A／D変換IC907によりサンプリングされた16ビットの画像データに対して、シリアルデータからパラレルデータにシリアル／パラレル変換を行ってレジスタ等に一旦格納しておくためのものである。

【0094】909は、画像取り込み部908により取り込んだ画像データ、例えば、会話音声の1フレーズ分の連続画像データを記憶しておくためのメモリであり、また、連続画像データを書き込めるだけの必要最小限の容量を持つメモリである。連続画像データのメモリへの書き込みは、CPU等のソフトウエア処理で行っても、専用のハードウエアで行っても良い。

【0095】910は画像認識処理部であり、メモリ909に書き込まれた連続画像データに対して、デジタルフィルタ，画像変換，2値化処理，画像解析，特徴抽出，照合，判定等の一連の画像認識処理を行う。ここで、画像認識に必要な画像モデルデータ，辞書データ，文法データは、画像認識処理部105内において、メモリ等に登録して格納しておく。画像認識処理は、CPUやDSP等のソフトウエア処理で行っても、専用のハードウエアで行っても良い。ここで、画像認識処理した結果は、音声認識処理部に渡す。

【0096】911は、前記音声認識処理部905から出力された会話音声の認識結果に対して翻訳したい言語に翻訳処理を行う翻訳処理部である。音声認識処理部905から出力される認識結果は、例えば、日本語であれば名詞，助詞，動詞，副詞等のかな漢字まじりのテキスト文章である。翻訳処理では、これらのかな漢字まじり文章に対して、構文解析及び辞書，文法規則，事例等からの文章生成を行い、翻訳結果を出力する。

【0097】912は、前記翻訳処理部911から出力された翻訳結果に対して、会話文に適した音声に変換して音声出力する音声合成処理部である。この音声合成処理部912では、より自然な会話文音声にするために、文章を構成している単語の発音やアクセント、更に、文章全体の抑揚を最適化して会話文の音声合成を行い、相手側に対して聞き取りやすい自然な音声を出力するための処理も行う。

【0098】913は16ビットのデジタル／アナログ（D／A）変換ICであり、前記音声合成処理部912から出力された音声のデジタル信号を、例えば、ローパスフィルタ（LPF）を経由して音声周波数帯域12kHzでアナログ信号に変換する。

【0099】914は、音声認識結果の途中経過や翻訳結果をテキストで表示するための液晶ディスプレイ（LCD）である。

18

【0100】915は、音声認識結果やその途中経過，翻訳結果を音声合成して音声出力するためのスピーカである。

【0101】916は、前記音声認識処理部905から出力された音声及び画像の認識結果に対して、誤認識部分の修正，補正を行う認識結果修正部である。誤認識部分を含んだ認識結果を、翻訳処理部911で翻訳すべき会話文の認識結果として、ただちに転送して翻訳させると、誤った翻訳結果になってしまう。そこで、認識結果修正部916は、認識結果を翻訳する前に、音声会話文を自分側で翻訳したい適切な文に修正することで翻訳精度を高めることができるようにする。

【0102】

【発明の効果】以上のように、本発明によれば、国際化時代における翻訳精度の高い音声認識翻訳装置の実現や、海外旅行先等で少しでも会話らしい相互のコミュニケーションをアシストする翻訳精度の高い携帯型音声認識翻訳機の実現が可能となる。

【図面の簡単な説明】

【図1】本発明になる音声画像認識翻訳装置の一形態形態を示すブロック図である。

【図2】図1に示した本発明になる音声画像認識装置における画像取り込み及び画像処理を示す説明図である。

【図3】図1に示した本発明になる音声画像認識装置の画像取り込み用カメラ内蔵マイクの構成図である。

【図4】図1に示した本発明になる音声画像認識装置における音声及び画像取り込み方法及び音声画像認識翻訳を示す説明図である。

【図5】図1に示した本発明になる音声画像認識装置の音声及び画像取り込み方法を示す説明図である。

【図6】図1に示した本発明になる音声画像認識装置における音声及び画像取り込み方法を示すフローチャート図である。

【図7】図1に示した本発明になる音声画像認識装置における音声データの判断方法を示す説明図である。

【図8】図1に示した本発明になる音声画像認識装置を携帯型翻訳機に適用した一例を示す説明図である。

【図9】本発明になる音声画像認識翻訳装置の他の実施形態を示すブロック図である。

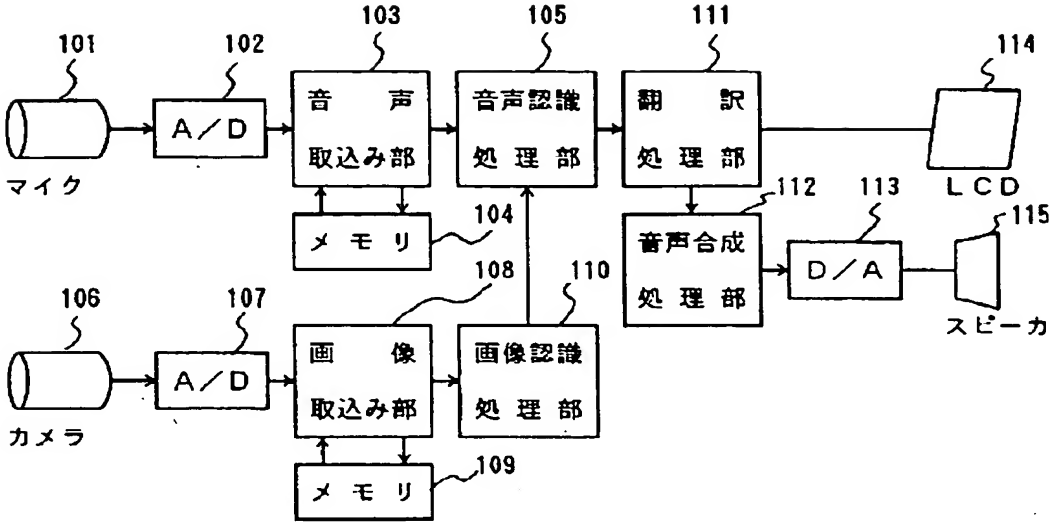
【図10】従来の携帯型の音声翻訳装置の構成を示すブロック図である。

【符号の説明】

101…マイク、102…A／D変換IC、103…音声取り込み部、104…メモリ、105…音声認識処理部、106…カメラ、107…A／D変換IC、108…画像取り込み部、109…メモリ、110…画像認識処理部、111…翻訳処理部、112…音声合成処理部、113…D／A変換IC、114…LCD、115…スピーカ。

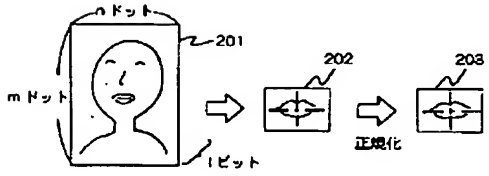
【図1】

図 1



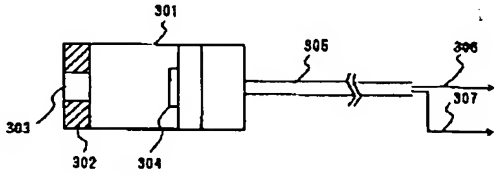
【図2】

図 2



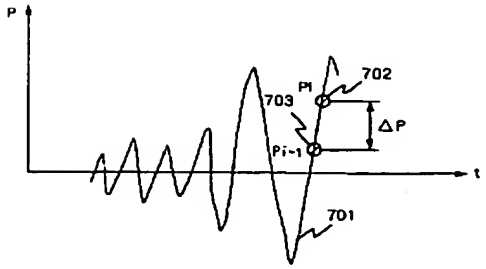
【図3】

図 3



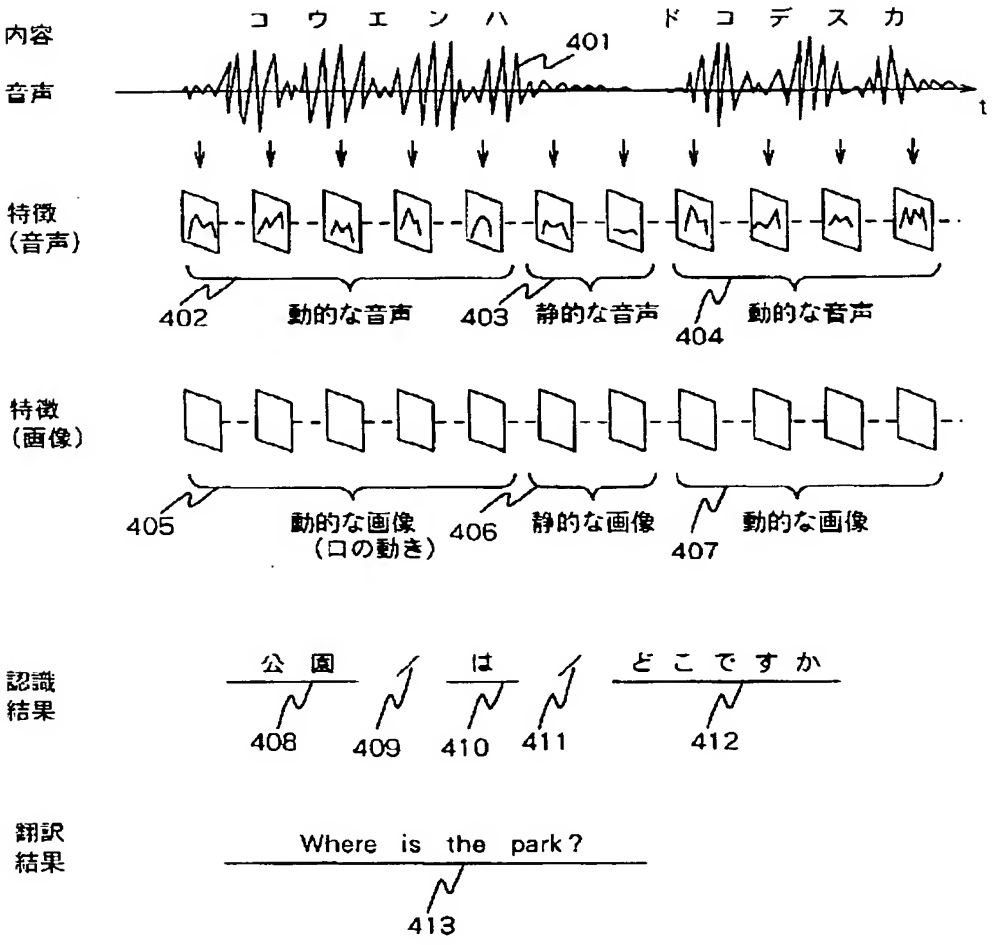
【図7】

図 7



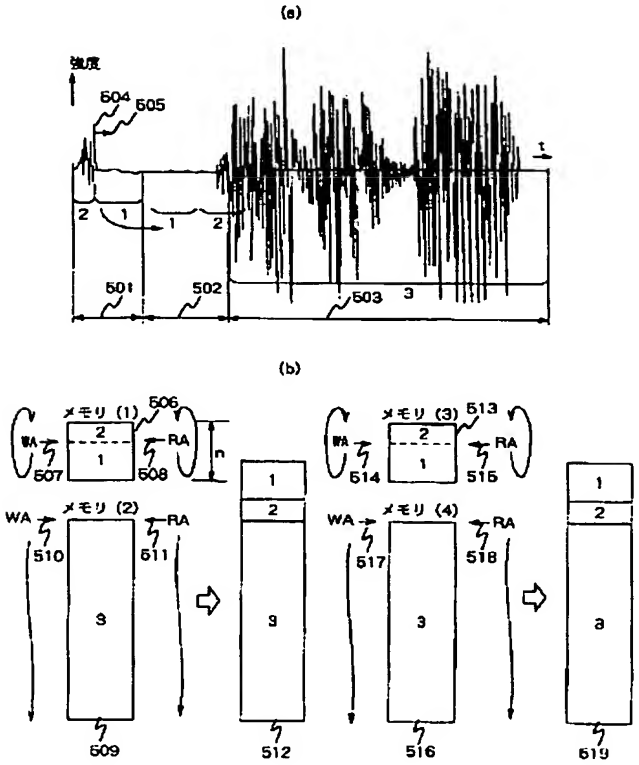
【図4】

図 4


内容
コ ウ エ ン ハ ド コ デ ス カ
401
音声
t
特徴
(音声)
402
動的な音声
403
静的な音声
404
動的な音声
特徴
(画像)
405
動的な画像
(口の動き)
406
静的な画像
407
動的な画像
認識
結果
公 園
は
どこですか
408
409
410
411
412
翻訳
結果
Where is the park?
413

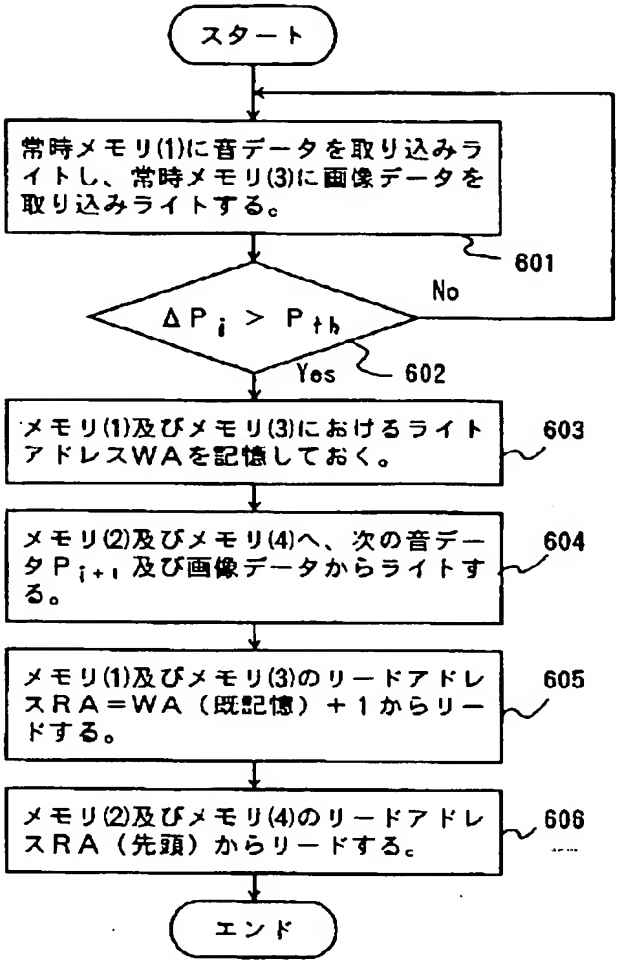
【図5】

図 5



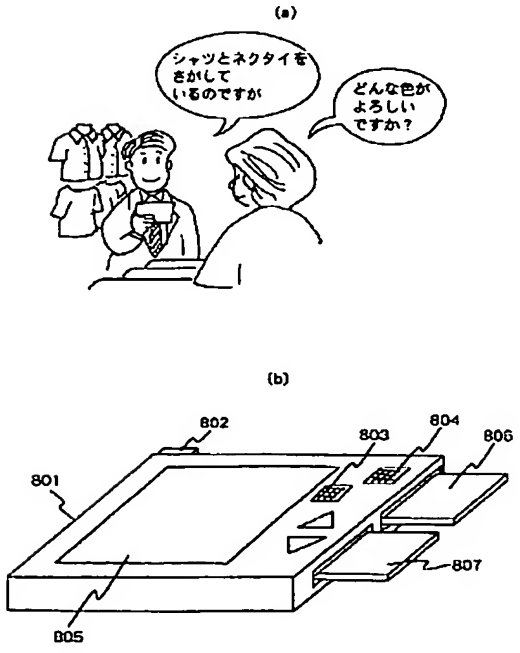
【図6】

図 6



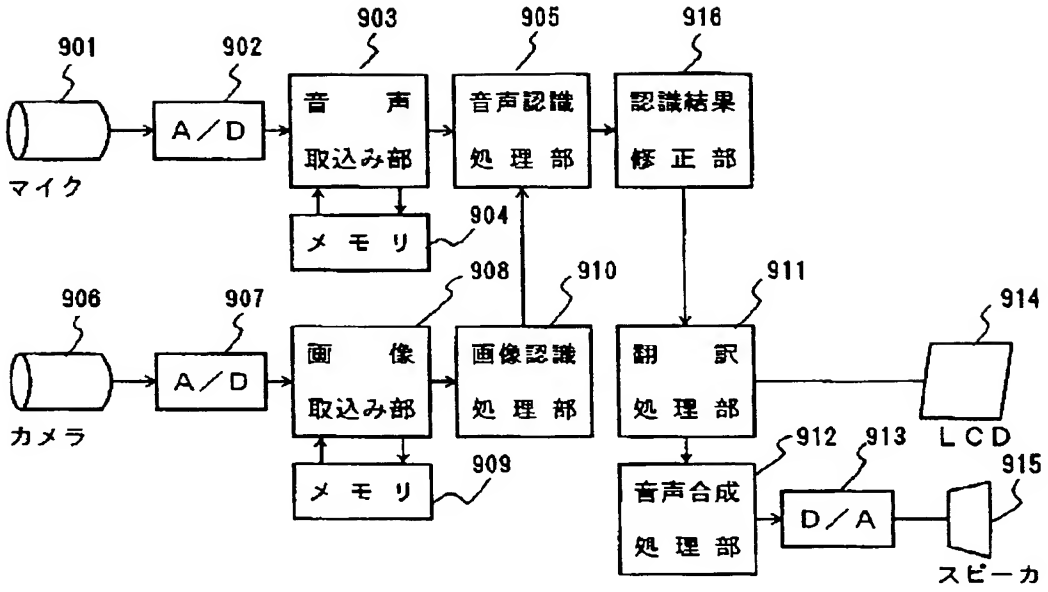
【図8】

図 8



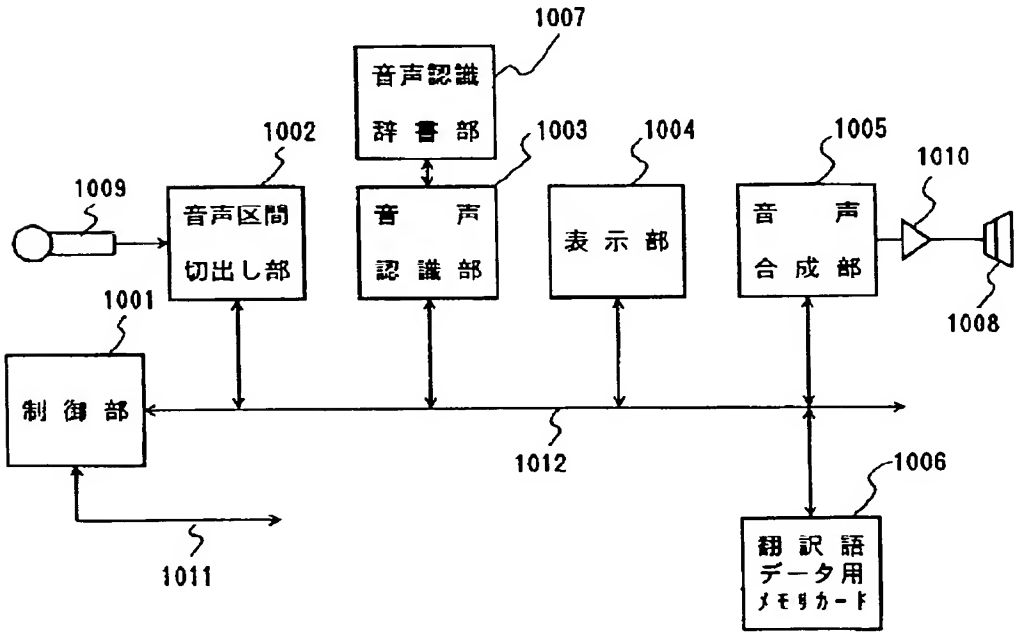
【図9】

図 9


901
902
903
905
916
マイク
A／D
音 声 取込み部
音声認識 処 理 部
認識結果 修 正 部
メ モ リ
904
906
907
908
910
911
914
カメラ
A／D
画 像 取込み部
画像認識 処 理 部
翻 訳 処 理 部
LCD
メ モ リ
909
912
913
915
音声合成 処 理 部
D／A
スピーカ

【図10】

図 10


1007
音声認識 辞 書 部
1002
1003
1004
1005
1010
1009
音声区間 切出し部
音 声 認 識 部
表 示 部
音 声 合 成 部
1001
1008
制 御 部
1012
1006
1011
翻 訳 語 データ用 メモリカード

フロントページの続き

(51)Int.Cl.[6]	識別記号	FI
		G06F 15/35　A

(72)発明者　天野　明雄
東京都国分寺市東恋ケ窪一丁目280番地
株式会社日立製作所中央研究所内
(72)発明者　伊東　功二
東京都小平市上水本町五丁目20番1号　株式会社日立製作所半導体事業部内
(72)発明者　佐藤　裕子
東京都小平市上水本町五丁目20番1号　株式会社日立製作所半導体事業部内
(72)発明者　石渡　一嘉
東京都小平市上水本町五丁目20番1号　株式会社日立製作所半導体事業部内